Combining Statistical Parameteric Speech Synthesis and Unit-Selection for Automatic Voice Cloning

Matthew P. Aylett and Junichi Yamagishi. Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning. In Proc. LangTech 2008, Brisbane, Australia, September 2008.

Combining Statistical Parameteric Speech Synthesis and Unit-Selection for Automatic Voice Cloning

Matthew P. Aylett[1,2], *Junichi Yamagishi*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, U.K.
[2]Cereproc Ltd., U.K.

matthewa@cereproc.com jyamagis@inf.ed.ac.uk

Abstract

The ability to use the recorded audio of a subject's voice to produce an open-domain synthesis system has generated much interest both in academic research and in commercial speech technology. The ability to produce synthetic versions of a subjects voice has potential commercial applications, such as virtual celebrity actors, or potential clinical applications, such as offering a synthetic replacement voice in the case of a laryngectomy. Recent developments in HMM-based speech synthesis have shown it is possible to produce synthetic voices from quite small amounts of speech data. However, mimicking the depth and variation of a speaker's prosody as well as synthesising natural voice quality is still a challenging research problem. In contrast, unit-selection systems have shown it is possible to strongly retain the character of the voice but only with sufficient original source material. Often this runs into hours and may require significant manual checking and labelling.

In this paper we will present two state of the art systems, an HMM based system HTS-2007, developed by CSTR and Nagoya Institute Technology, and a commercial unit-selection system CereVoice, developed by Cereproc. Both systems have been used to mimic the voice of George W. Bush (43rd president of the United States) using freely available audio from the web. In addition we will present a hybrid system which combines both technologies. We demonstrate examples of synthetic voices created from 10, 40 and 210 minutes of randomly selected speech. We will then discuss the underlying problems associated with voice cloning using found audio, and the scalability of our solution.

Index Terms: speech synthesis, unit-selection, statistical parametric synthesis, voice cloning, HMM, speaker adaptation

1. Introduction

Vocal mimicry by computers is regarded with both awe and suspicion [1]. This is partly because perfect vocal mimicry is also the mimicry of our own sense of individuality: the use of a certain voice draws with it much more than the voice itself, it also draws the associations we have with that voice. Conveying this sense of character is becoming important in a whole set of innovative applications for human-computer interfaces which use speech for input and output.

For example, the ability to produce synthetic versions of a subjects voice has potential attractive commercial applications, such as virtual celebrity actors, or potential beneficial clinical applications, such as offering a synthetic replacement voice in the case of a laryngectomy. In addition, the ability to retain the character of a speaker could be combined with translation systems, where it would help personalize *speech-to-speech* translation so that a user's speech in one language can be used to produce corresponding speech in another language while continuing to sound like the user's voice. It might eliminate the need for subtitles and onerous voice-overs acting on international broadcasts or movies in the future.

In this paper we investigate the reproduction/mimicry aspects of up-to-date speech synthesis technologies: how well can we take a well-known speaker and duplicate his acoustic feature, linguistic features, and speaking styles so that a listener immediately recognises the speaker? Furthermore, how effective is this mimicry for conveying the character of the speaker in an amusing manner? We term the process of producing a speech synthesis system that can effectively mimic a speak "voice cloning". We apply two major competing technologies to this voice cloning problem, the first is a well-established and well-studied technique called "unit-selection", which concatenates segments of speakers' source speech to create new utterances [2], the second is often termed "statistical parametric synthesis," where a statistical acoustic model is trained or adapted from speakers' source speech [3]. In the experiments, we will apply both techniques to the problem of cloning the voice of **George W. Bush** (The 43rd President of the United States) and produce a short rendition of the introduction of a well known children's story, "The Emperor's New Clothes". In addition we will explore the use of a new hybrid system which attempts to utilise the strengths of both approaches to create a more scalable means of mimicking voices.

2. Voice Cloning

2.1. Constraints

There is a genuine commercial interest in voice cloning for entertainment as well as an interest for speakers to create a virtual version of there own voice for use in cyber-realities varying from web pages to virtual environments. In addition there is a serious clinical application for the technology where it can be used to produce synthetic voices for patients who, due to illness, trauma, or surgery can no longer speak normally. However there are three constraints which have made voice cloning a rare activity in speech synthesis.

1. The resulting synthesis must sound 'natural' enough to effectively mimic the voice. Only over the last few years has speech synthesis begun to reach this level of naturalness.
2. The amount of data required from a speaker is not unlimited. Ideally we wish to mimic voices from as small amount of material as possible.
3. The type of speech styles required have a big impact on

Table 1: *Speech synthesis systems under test. These mnemonics will be used throughout this paper to refer to specific voices.*

System	Source Data		
	10 minutes	40 minutes	210 minutes
HTS-2007	HTS10	HTS40	HTS210
Cereproc CereVoice	CPCV10	CPCV40	CPCV210
Cereproc Hybrid	CPHY10	CPHY40	CPHY210

current cloning techniques. To a very large extent we can successfully mimic a voice in a single speech style with between 3-5 hours of carefully recorded speech. However to mimic a voice across many speech styles, for example mimicking different emotions is still a challenging research problem.

The unit-selection techniques can produce good mimicry of a single speech style (and recently some speech style variation [4]) given sufficient carefully collected data. The statistical parametric approaches, while not reaching the same level of naturalness as that of the unit-selection techniques, can offer the ability to mimic voices with substantially smaller amount of source speech data. In addition the statistical parametric techniques make it easier to alter vocal style due to the model based approach.

In this paper we will present three systems — HTS-2007 (statistical parametric approach) [5][6], Cereproc CereVoice (unit-selection approach) [4], and Cereproc Hybrid (hybrid approach of statistical parametric and unit-selection) — based on three different amounts of source speech material, a 10 minute database, a 40 minute database, and a 210 minute database taken from audio publicly available of Mr. George W. Bush.

2.2. Data Collection

The source data for these research voices was taken from audio freely available on the web. In order to effectively create the voices audio was carefully chosen and segmented into utterances varying from 1-261 words in length (Mean 12 SD 8.08). Note that care was taken to avoid background noise (i.e. applause, music), disfluencies (Ums and ahs), and poor audio quality caused by compression or low sampling rates. The data was manually transcribed and then verified using Cereproc's proprietary voice building system.

From 257 minutes of source speech in 4006 speech utterance files obtained via the above procedures, three sets of utterance lists were randomly selected for generating voice using approximately 10, 40 and 210 minutes of speech. For all the systems only this amount of source material was then used to generate the resulting voice. For example acoustic models were not trained on all the data and then used to segment part of it. From these three lists nine voices were created for each system, statistical parametric, unit-selection, and hybrid systems. See Table 1 for the mnemonics used for each voice.

3. Statistical Parametric Synthesis: HTS-2007

The HTS-2007 system is a high-quality speaker-independent HMM-based speech synthesis system developed by CSTR and

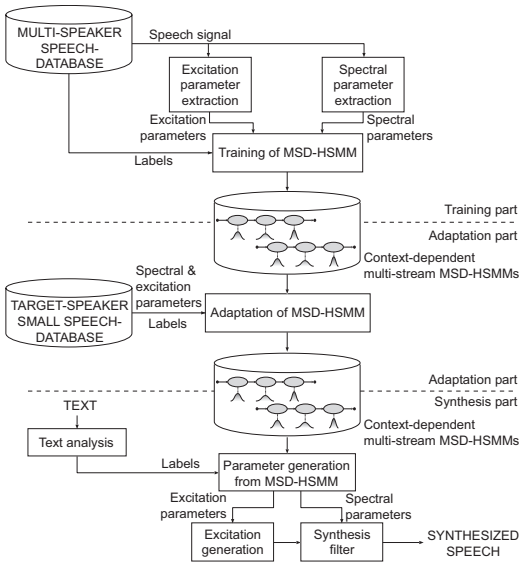

Figure 1: Overview of the HTS-2007 speech synthesis system.

Nagoya Institute Technology. In the system (Fig. 1), an average voice model using context-dependent multi-stream MSD-HSMMs is created from more than 10 hours of speech data uttered by many speakers and is adapted with speech data obtained from a target speaker. The acoustic features for the MSD-HSMMs are three kinds of parameters required for a high-quality speech vocoding method with mixed-band excitation called *STRAIGHT* [6]: the STRAIGHT mel-cepstrum, $\log F_0$, and aperiodicity measures. Using the above acoustic features, the MSD-HSMMs are trained based on the speaker-adaptive training and are adapted to the target speaker by using a combined algorithm of constrained structural maximum a posteriori linear regression (CSMAPLR) [7] and maximum a posteriori (MAP) adaptation. Speech parameters are directly generated from the adapted MSD-HSMMs using a penalised maximum likelihood method [8].

Since the average voice models can utilise the large-scale speech database and both spectral and prosodic features such as $\log F_0$ or phone duration can be statistically and simultaneously transformed from the average voice model into those of the target speaker, we can robustly create voices even from relatively small amount of speech data. However, the synthetic speech generated from the voices has a "buzzy" quality, since speech waveform is vocoded from pulse or noise excitation. Parts of the system have already been released in an open-source software toolkit called HTS (from "H Triple S," an initialism for the "HMM-based speech synthesis system") [9].

4. Unit-Selection Synthesis: Cereproc CereVoice

CereVoice is a faster-than-realtime diphone unit selection speech synthesis engine, available for academic and commercial use. The core CereVoice engine is an enhanced synthesis 'back end', written in C for portability to a variety of platforms. The engine does not fit the classical definition of a synthesis back end, as it includes lexicon lookup and letter-to-sound rule modules, see Fig. 2. An XML API defines the input to the en-

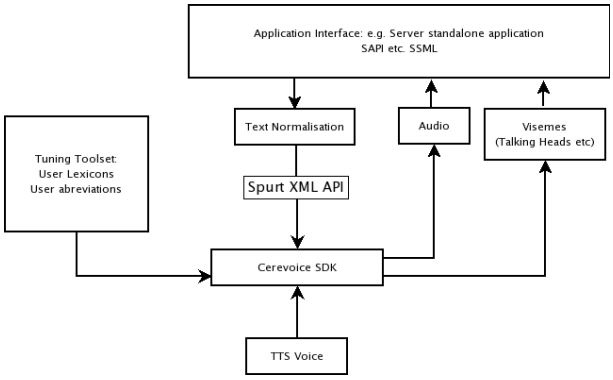

Figure 2: *Overview of the architecture of the CereVoice synthesis system. A key element in the architecture is the separation of text normalisation from the selection part of the system and the use of an XML API.*

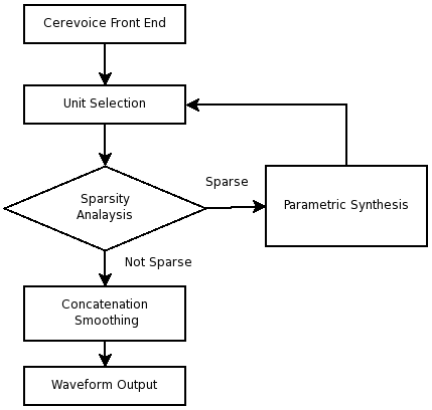

Figure 3: *Combining parametric and unit selection synthesis.*

gine. The API is based on the principle of a 'spurt' of speech. A spurt is defined as a portion of speech between two pauses. To simplify the creation of applications based on CereVoice, the core engine is wrapped in higher level languages such as Python using Swig. For example, a simple Python/Tk GUI was written to generate the test sentences for the Blizzard challenge.

The CereVoice engine is agnostic about the 'front end' used to generate spurt XML. CereProc use a modular Python system for text processing. Spurt generation is carried out using a greedy incremental text normaliser. Spurts are subsequently marked up by reduction and homograph taggers to inform the engine of the correct lexical variant dependent on the spurt context.

5. Hybrid Approach: Cereproc Hybrid

A key weakness in the unit selection approach is the issue of sparsity. In order to produce a smooth rendition of speech the database must contain appropriate units. If these are diphones and are American voice contains 40 phones this would require in 1600 different units for full coverage. In addition to the phones, you also need coverage of prosodic context, for example stress, increasing the required units to 6400 units if you include phrasing 25.6k units for full coverage. Finally the context of many units is also vital for concatenation because of coarticulation. If you also require coverage of all left and right contexts you require over 4 million different units.

Fortunately many of these contexts are very rare or do not occur. However, even with a modest requirement of 1600 different diphones, because speakers are required (in general) to produce normal connected speech for the source database, this tends to result in a database of approximately 300k diphones which can require up to 21 hours of studio recording time. Even given this size of database there will be many contexts missing and this in turn can produce concatenation errors.

Parametric approaches offer an attractive solution to this sparsity problem. Firstly, as the speech is synthesised from model parameters there are no concatenation errors. Secondly because the voice is derived from a model it is possible to use adaptation to harness information from other speakers to improve the model on only a small selection of data. The disadvantages with the parametric approach is that the generation of the speech from model parameters does not produce completely natural sounding speech. In some cases a so called 'vocoder buzz' is perceivable. However a more profound problem is that it is necessary to model **all** features in speech including prosodic variation and structure. Natural speech prosody is complex and as yet not entirely understood. Thus parametric systems can have dull or repetitive prosody.

A hybrid approach tries to use the advantages of both systems to create a more saleable and natural solution. Cereproc have developed a means for seamlessly concatenating parametric produced speech within a unit selection framework. In effect, when sparsity of concatenation errors are assessed as likely, sections of parametric speech can be used rather than standard units, see Fig. 3. The advantage of this approach is that a large proportion of the prosody can be produced within unit selection while at the same time avoiding sparsity caused by high dimensionality.

Results in this paper are for a very early prototype system. The system is combination of Cereproc CereVoice combined with HTS-2007.

6. Evaluation

All the systems were used to synthesise the opening paragraph of "The Emperor's New Clothes" by Hans Christian Anderson. This text was used for evaluation because: the material was completely different from the domain of the source material, story telling is harsh test of prosodic relevance and variation, the vocabulary was simple which meant that intelligibility was less likely to be a factor in the assessment.

The paragraph was split up into 9 utterances varying from 8 to 31 words long. Each subject heard each utterance from each of the systems and scored the naturalness on a five point scale. Finally the full paragraph from one system was played to the subject who then gave an overall score for the complete rendition from that specific system. 23 subjects took part in the experiment (of which 9 were native speakers).

Fig. 4 shows a histogram of the average mean opinion score (MOS) for each system. 95% confidence intervals are shown for each histogram.

The full audio for each system are available on the web at `http://www.cogsci.ed.ac.uk/~matthewa/LANGSPEECH2008.htm`.

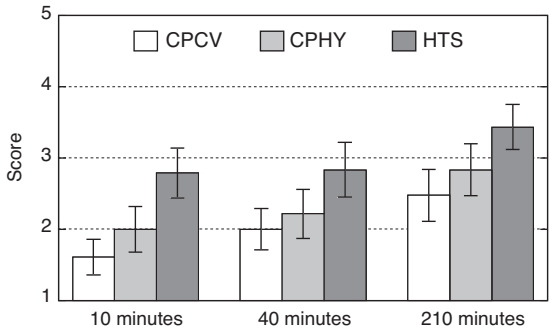

Figure 4: *Average 5 point MOS scores for all systems and databases. Error bars show 95% confident intervals.*

7. Discussion

Due to resource limitations our evaluation was small and the results should be treated with caution. In addition, MOS style evaluations can be problematic because there can be significant subject variation in what is regarded as "natural". For example non native speakers rated all systems approximately 0.5 worse on the MOS scale than native speakers.

There is also, arguably, a tendency for concatenation errors to be more heavily penalised by subjects than the problem of vocoder buzz when evaluated in single sentence MOS experiments.

However there is a clear preference for the HTS system in contrast to the past comparison of speech synthesis systems (e.g. [10]) in which hybrid approaches provide significant better quality. This in part may be related to specific issues with regards to found data (as opposed to carefully recorded data). In general one of the strengths of unit selection is that you get a lot of speech which has had very little modification made to it. When the speech is carefully recorded in a quite environment this is a great advantage. With this data, however, much of the audio was recorded in very different environments (in some cases decompressed from MP3). One of the strengths of the parametric approach was that it was able to remove the inconsistency caused by recording environment. Another important problem caused by recording environment is the possibility of phase differences between different sections of audio. Such phase variation can make time domain concatenation extremely problematic. Phase distortion was such a problem that the Cereproc system for smoothing the vocoded speech had to be switched off.

Finally we intentionally did not choose any data on the basis of it improving unit coverage. Thus the results highlight another of the strengths of parametric synthesis, where, for the 10 minutes database, the unit selection system was completely unable to function effectively in complete contrast to the Hybrid and HTS systems.

8. Conclusions

Given the data available we feel that the results for relatively small databases were excellent for the HTS system, which maintained impressive consistency. The Hybrid system exhibited teething problem in terms of effectively merging different recording environments but again showed that a dual approach is a serious research direction in speech synthesis. For the large database, all systems performed well, with HTS doing best in terms of a sentence by sentence 5 point MOS evaluation.

However we strongly suggest interested readers listen to the 9 short versions of the audio themselves to gain an insight into the differences between these systems and pros and cons of them. Since the artificial and buzzy quality of the synthetic speech generated from the HTS system remains, we need to explore a better hybrid algorithm which can work robustly and effectivley, even for found data, in order to reproduce the depth and variation of a speaker's prosody.

9. Acknowledgements

The authors would like to thank the HTS working staffs. Support for this research was provided by EPSRC (award number EP/D058139/1).

10. References

[1] L. Seward, "Scientists warn of 'vocal terror'," BBC NEWS, Sept. 2007. http://news.bbc.co.uk/1/hi/sci/tech/6994595.stm

[2] A. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", Proc. ICASSP-96, pp.373–376, May 1996.

[3] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," Proc. ICASSP 2007, pp.1229–1232, April 2007.

[4] M.P. Aylett, C.P. Pidcock, "The CereVoice characterful speech synthesiser SDK," Proc. AISB 2007, pp.174–178, April, 2007

[5] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007," Proc. BLZ3-2007 (in Proc. SSW6), Aug. 2007.

[6] J. Yamagishi, T. Nose, H. Zen, T. Toda, K. Tokuda, S. King, and S. Renals, "A speaker-independent HMM-based speech synthesis system for the Blizzard Challenge 2007," IEEE Trans. Speech, Audio & Language Process., 2007 (under review).

[6] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds", Speech Communication, vol. 27, pp.187–207, 1999.

[7] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR Adaptation Algorithm", IEEE Trans. Speech, Audio & Language Process., 2007 (under review).

[8] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis", IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.816–824, May 2007.

[9] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.B. Black and T. Nose, "The HMM-based speech synthesis system (HTS) Version 2.0.1", 2007. http://hts.sp.nitech.ac.jp/

[10] R.A.J Clark, M. Podsiadlo, M. Fraser, C. Mayo, S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results", http://festvox.org/blizzard/bc2007/